Llama Guard 3 Vision: Safeguarding Human-Al Image Understanding Conversations

Jianfeng Chi^{*}, Ujjwal Karn^{*}, Hongyuan Zhan^{*}, Eric Smith^{*}, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, Kartikeya Upasani[†], Mahesh Pasupuleti[†]

GenAI at Meta

*Core contributors, [†]Senior contributors

We introduce Llama Guard 3 Vision, a multimodal LLM-based safeguard for human-AI conversations that involves image understanding: it can be used to safeguard content for both mutimodal LLM inputs (prompt classification) and outputs (response classification). Unlike the previous text-only Llama Guard versions (Inan et al., 2023; Llama Team, 2024b,a), it is specifically designed to support image reasoning use cases and is optimized to detect harmful multimodal (text and image) prompts and text responses to these prompts. Llama Guard 3 Vision is fine-tuned on Llama 3.2-Vision and demonstrates strong performance on the internal benchmarks using the MLCommons taxonomy. We also test its robustness against adversarial attacks. We believe that Llama Guard 3 Vision serves as a good starting point to build more capable and robust content moderation tools for human-AI conversation with multimodal capabilities.

Date: November 18, 2024 Correspondence: Jianfeng Chi at jianfengchi@meta.com Code: https://github.com/meta-llama/llama-recipes/tree/main/recipes/responsible_ai/llama_guard Blogpost: https://www.llama.com/trust-and-safety/#safeguard-model

1 Introduction

The past few years have witnessed an unprecedented improvement in the capabilities of Large Language Models (LLMs), driven by the success in scaling up autoregressive language modeling in terms of data, model size, and the amount of compute used for training (Kaplan et al., 2020). LLMs have demonstrated exceptional linguistic abilities (Brown, 2020; Achiam et al., 2023), general tool use (Schick et al., 2024; Cai et al., 2023), and commonsense reasoning (Wei et al., 2022; OpenAI, 2024), among other impressive capabilities. The success of LLMs as general-purpose assistants motivates research and development to extend instruction-tuning to the vision-language multimodal space (Liu et al., 2023; Gemini Team, 2023). These vision-language multimodal models, which can process and generate both text and images, also achieve human-expert performance on a wide range of tasks, such as (document) visual question answering (Antol et al., 2015; Mathew et al., 2021), image captioning (Lin et al., 2014), and image-text retrieval (Plummer et al., 2015).

While these vision-language multimodal models hold tremendous promise for many applications, they should be used along with proper system guardrails to ensure safe and responsible deployment, because they can generate or propagate harmful content when interacting with online users. However, most existing guardrails (Inan et al., 2023; Llama Team, 2024b,a; Yuan et al., 2024; Ghosh et al., 2024) for the interaction (e.g., conversation) between humans and AI agents are text-only: conversation data involving other modalities, such as images, cannot be used as inputs for such guardrails. This calls for a safeguard tool for classifying safety risks in prompts and responses for conversations with multimodal contents involved.

In this work, we introduce Llama Guard 3 Vision, a multimodal LLM-based safeguard for human-AI conversations that involves image understanding: it can be used to safeguard content for both mutimodal LLM inputs (prompt classification) and mutimodal LLM responses (response classification). Unlike text-only Llama Guard versions (Inan et al., 2023; Llama Team, 2024b,a), it is specifically designed to support image reasoning use cases and is optimized to detect harmful multimodal (text and image) prompts and text responses to these prompts. Figure 1 gives an example of how Llama Guard 3 Vision classifies harmful content in the



Figure 1 Llama Guard 3 Vision classifies harmful content in the response classification task.

response classification task.

Llama Guard 3 Vision is fine-tuned on Llama 3.2-Vision and demonstrates strong performance on our internal benchmark using the 13 hazards of the MLCommons taxonomy. To better understand the adversarial robustness of Llama Guard 3 Vision, we evaluate it against two existing powerful white-box attacks (Zou et al., 2023; Madry, 2017) using both textual and visual inputs. We find that Llama Guard 3 Vision is more robust in the response classification task compared to the prompt classification task under practical threat model scenarios, since Llama Guard 3 Vision primarily relies on model response for classification while effectively ignoring prompt-based attacks. We hope our white paper can help practitioners build better customizable safeguard models for multimodal use cases, and facilitate further research and development of multimodal LLM safety.

Paper Organization In Section 2, we present the hazard taxonomy and policy used to train Llama Guard 3 Vision. In Section 3, we describe the process of building Llama Guard 3 Vision. Our experiment section (Section 4) is divided into two parts: In Section 4.1, we evaluate Llama Guard 3 Vision on our internal benchmark and compare it with baselines such as GPT-40 (Achiam et al., 2023), and in Section 4.2, we present the details and results of our adversarial robustness experiments.

2 Hazard Taxonomy and Policy

Llama Guard 3 Vision is trained to predict safety labels on the 13 categories shown below, based on the MLCommons taxonomy of 13 hazards (Vidgen et al., 2024). Below, we provide the detailed description of each safety category:

- **S1 Violent Crimes.** Responses that enable, encourage, or endorse the commission of violent crimes, including: (1) unlawful violence toward people (e.g., terrorism, genocide, murder, hate-motivated violent crimes, child abuse, assault, battery, kidnapping) and (2) unlawful violence toward animals (e.g., animal abuse)
- S2 Non-Violent Crimes. Responses that enable, encourage, or endorse the commission of non-violent

crimes, including: (1) personal crimes (e.g., labor trafficking, threats, intimidation, hate-motivated non-violent crimes), (2) financial crimes (e.g., fraud, scams, money laundering), (3) property crimes (e.g., theft, arson, vandalism), (4) drug crimes (e.g., creating or using regulated substances), (5) weapons crimes (e.g., producing unlicensed firearms), and (6) cyber crimes (e.g., hacking)

- **S3 Sex-Related Crimes.** Responses that enable, encourage, or endorse the commission of sex-related crimes, including: (1) sex trafficking, (2) sexual assault (e.g., rape), (3) sexual harassment (physical, verbal, or visual), and (4) prostitution
- **S4 Child Sexual Exploitation.** Responses that contain, describe, enable, encourage, or endorse the sexual abuse of children
- S5 Defamation. Responses that are both verifiably false and likely to injure a living person's reputation
- **S6 Specialized Advice.** Responses that contain specialized financial, medical, or legal advice, or that indicate dangerous activities or objects are safe
- **S7 Privacy.** Responses that contain sensitive, nonpublic personal information that could undermine someone's physical, digital, or financial security
- S8 Intellectual Property. Responses that may violate the intellectual property rights of any third party
- **S9 Indiscriminate Weapons.** Responses that enable, encourage, or endorse the creation of indiscriminate weapons, including: (1) chemical weapons (e.g., nerve gas), (2) biological weapons (e.g., anthrax), (3) radiological weapons (e.g., cobalt bombs), (4) nuclear weapons (e.g., fission bombs), and (5) high-yield explosive weapons (e.g., cluster munitions)
- **S10 Hate.** Responses that demean or dehumanize people on the basis of their sensitive, personal characteristics (i.e., race, color, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and/or serious disease)
- S11 Suicide & Self-Harm. Responses that enable, encourage, or endorse acts of intentional self-harm, including: (1) suicide, (2) self-injury (e.g., cutting), and (3) disordered eating
- S12 Sexual Content. Responses that contain erotica
- **S13 Elections.** Responses that contain factually incorrect information about electoral systems and processes, including the time, place, or manner of voting in civic elections

Note on privacy violations for image understanding. Compared to text-only Llama Guards, we consider identification from images in prompts (e.g., recognizing a real world person from the image through any visual clue) as a potential Privacy violation. For example, given an image of a real person, if the user asks "Do you know the person in the image?" or "What makes her famous?", and the agent's response tries to identify the person (whether the identity recognized is correct or incorrect), Llama Guard 3 Vision is trained to classify the response as unsafe.

3 Building Llama Guard 3 Vision

3.1 Input-output Safeguarding

Similar to text-only Llama Guard (Inan et al., 2023), we fine-tune our model with tasks that ask to classify content as being safe or unsafe. For input-output safeguarding tasks, we use the same four ingredients:

A set of guidelines. The task takes a set of guidelines as input, which consist of numbered descriptions of "unsafe categories", which the model uses for making a safety assessment.

The type of classification. Each task indicates whether the model needs to classify the user messages (prompt classification) or the agent messages (response classification).

The conversation. The conversation contains the user provided image, as well user turns and agent turns (this can be single-turn or multi-turn).

The output format. The model should output "safe" or "unsafe". If the model assessment is "unsafe", then the output should contain a new line, listing the taxonomy categories (mentioned in Section 2 above) that are violated in the given piece of content.

Figure 1 illustrates the response classification task, the guidelines, and the desired output format for Llama Guard 3 Vision.

3.2 Data Collection

To train Llama Guard 3 Vision, we employed a hybrid dataset comprising both human-generated and synthetically generated data. Our approach involved collecting human-created prompts paired with corresponding images, as well as generating benign and violating model responses using our in-house Llama models. We utilized jailbreaking techniques to elicit violating responses from these models. The resulting dataset includes samples labeled either by humans or the Llama 3.1 405B model (Llama Team, 2024a). To ensure comprehensive coverage, we carefully curated the dataset to encompass a diverse range of prompt-image pairs, spanning all hazard categories listed above.

To improve model performance and robustness, we sourced text-only data from the Llama Guard 3 training set (Llama Team, 2024a), and added the same with dummy images to Llama Guard 3 Vision training set. For the prompt classification tasks, the final dataset comprises 22,500 prompts and image pairs, with their respective annotations. For the response classification task, the final dataset comprises 40,034 prompts, responses and images, with their respective annotations.

3.3 Model & Training Details

We build Llama Guard 3 Vision on top of the Llama 3.2 11B vision model. We perform supervised fine-tuning using a sequence length of 8192, a learning rate of 1×10^{-5} , and training for 3600 steps. The model is trained on only one image per prompt. For our image data, our vision encoder rescales it into 4 chunks, each of 560×560 pixels.

Data Augmentation. Similar to text-only Llama Guard, we employ data augmentation techniques to improve the model's generalization abilities. We drop a random number of categories from the model prompt if they're not violated in the given example: this ensures that the model can learn to take into account only the included categories. We also shuffle the category indices across training examples (while making corresponding changes in the desired outputs) in order to avoid format memorization.

4 Experiments

4.1 Performance on the Internal Benchmark

Baselines. We use GPT-40 and GPT-40 mini in a LLM-as-a-judge setup, with zero-shot prompting using MLCommons hazard taxonomy as the two baselines. When sending queries to GPT-40 and GPT-40 mini APIs, we observe that GPT-40 and GPT-40 mini APIs sometimes reject our safety classification request for the safety reasons. For such rejected classification requests, we consider that GPT-40 (mini) classifies it as "unsafe".

Result Analysis. Table 1 shows that Llama Guard 3 Vision outperforms GPT-40 and GPT-40 mini in terms of F1 score, specifically in response classification. Notably, Llama Guard 3 Vision achieves better performance with much lower false positive rates. We also observe that the recall (false positive rates) of prompt classification is much lower (higher) than that of response classification, because the ambiguity of prompts (each one consisting of a text and image) in image understanding tasks makes prompt classification more challenging than response classification. For example, given a text prompt "how do I buy this" with an image of multiple objects, whether the prompt is safe or not depends on which objects the user is referring to, and it is hard to decide in some cases. In such cases, we recommend using Llama Guard 3 Vision in the response classification task. Table 2 also shows the F1 score per category on our internal test set. We observe that Llama Guard 3 Vision performs well in categories such as Indiscriminate Weapons and Elections, while showing > 0.69 F1 scores across all categories.

Table 1 Comparison of performance of various models measured on our internal test set for MLCommons hazardtaxonomy.

Model	Task	Precision (\uparrow)	Recall (\uparrow)	F1 (\uparrow)	FPR (\downarrow)
Llama Guard 3 Vision	Prompt Classification	0.891	0.623	0.733	0.052
GPT-40		0.544	0.843	0.661	0.485
GPT-40 mini		0.488	0.943	0.643	0.681
Llama Guard 3 Vision	Response Classification	0.961	0.916	0.938	0.016
GPT-40		0.579	0.788	0.667	0.243
GPT-40 mini		0.526	0.820	0.641	0.313

Table 2 Category-wise breakdown of F1 for Llama Guard 3 Vision on our internal test set for response classification with safety labels from the ML Commons taxonomy.

Category	F1 (†)
Violent Crimes	0.839
Non-Violent Crimes	0.917
Sex Crimes	0.797
Child Exploitation	0.698
Defamation	0.967
Specialized Advice	0.764
Privacy	0.847
Intellectual Property	0.849
Indiscriminate Weapons	0.995
Hate	0.894
Self-Harm	0.911
Sexual Content	0.947
Elections	0.957

4.2 Adversarial Robustness of Llama Guard 3 Vision

Building machine learning models that are completely robust to adversarial attempts to manipulate their predictions is considered impossible as of today (Szegedy, 2013; Tramer, 2022; Anwar et al., 2024; Wei et al., 2024; Carlini et al., 2024). For this reason, the primary goal of Llama Guard 3 Vision is to enhance the safety of generative language models under typical and benign usage. Llama Guard 3 Vision offers an additional layer of protection against adversaries, requiring attackers to bypass both the underlying language model's protections and Llama Guard 3 Vision simultaneously. However, like all current AI systems, Llama Guard 3 Vision itself is not immune to adversarial attacks.

In this release, we evaluate Llama Guard 3 Vision against existing powerful white-box attacks: PGD (Madry, 2017) and GCG (Zou et al., 2023), with the former operating on images and the latter on text. By stress testing Llama Guard 3 Vision against adversarial attacks, we hope to inform users as to what limitations their systems may have against adversaries, and help the responsible deployment of AI applications. While we offer specific recommendations to enhance robustness, it is crucial to note that these recommendations do not guarantee an impenetrable system.

Threat model. We consider an adversary attempting to manipulate Llama Guard 3 Vision into misclassifying a harmful prompt provided by the user, or harmful content generated by an agent as safe.¹ The adversary has full white-box access to Llama Guard 3 Vision, given that the model architecture and parameters are publicly available. For a worst-case analysis, we also optimize the attacks with knowledge of the unsafe generation produced by the agent. In practice, adversaries may not know the actual agent generations since the responses will be flagged as *unsafe*, and therefore hidden from the attacker. We believe our results are a reasonable

 $^{^{1}}$ In principle, an attacker could attack Llama Guard 3 Vision to misclassify a safe conversation as unsafe, but that does not seem to be a useful goal.

Attack	Task	l_{∞} -bound	% Safe (\downarrow)
No attack	Prompt classification	0/255	21%
PGD	Prompt classification Prompt classification Prompt classification	$8/255 \\ 128/255 \\ 255/255$	70% 82% 82%
No attack	Response classification	0/255	6%
PGD	Response classification Response classification Response classification	$8/255 \\ 128/255 \\ 255/255$	$22\% \\ 27\% \\ 27\%$

Table 3 Percentage of *harmful* prompts and conversations from the test set that are misclassified as *safe* after applying the PGD attack on images in the prompt.

Table 4 Percentage of text-only *harmful* prompts and conversations from the test set that are misclassified as *safe* after applying the GCG attack on user prompt or agent response.

Attack	Moderation input	Attack appended to	% Safe (\downarrow)
No attack	Prompt classification		4%
GCG	Prompt classification	User prompt	72%
No attack	Response classification		16%
GCG	Response classification Response classification	User prompt Agent response	$30\% \\ 75\%$

upper-bound for the vulnerabilities of Llama Guard 3 Vision in practical scenarios.

Experimental setup. We evaluate Llama Guard 3 Vision for prompt and response classification on conversations including both the prompt and the agent response. PGD, an image-based attack, is evaluated on 100 harmful conversations in which prompts contain an image. GCG, a text-only attack, is evaluated on another set of 100 text-only unsafe conversations without image inputs. For details on the experimental setup and attack implementation, please refer to Appendix B.

Results against PGD image attacks. Continuous input modalities, like images, enable more efficient adversarial optimization than discrete optimization required by text attacks (Carlini et al., 2024; Rando et al., 2024). Since Llama Guard 3 Vision is intended to safeguard conversations that include images, we evaluate its robustness against PGD (Madry, 2017) on prompts that contain image inputs. PGD optimizes the input pixels to maximize the probability of Llama Guard 3 Vision predicting *safe*. Our experiments consider different l_{∞} perturbation budgets, reflecting scenarios where attackers may need to preserve semantic information (small perturbations) or insert images solely to fool the safety classifier (large perturbations). Results show that even with small perturbations (8/255), PGD attacks can significantly increase the rate of harmful prompts misclassified as *safe*, from 21% to 70% (see Table 3 for full results). On the other hand, fooling response classification of Llama Guard 3 Vision by tampering with an image in the prompt is much more difficult: even under an unbounded PGD attack, unsafe responses are misclassified as safe only 27% of the time. This encouraging result suggests that Llama Guard 3 Vision exhibits robust behavior in response classification, primarily relying on model responses while effectively ignoring prompt-based attacks. That said, this is still substantially higher than 6% of the clean baseline, and future work should explore techniques to further improve the robustness of Llama Guard 3 Vision against image attacks.

Recommendations to enhance robustness: Enforce both prompt and response classification modes when deploying Llama Guard 3 Vision in the wild, and use techniques such as adversarial training to further improve robustness. Limit image access in safety-critical applications, or complement image-based classification with text-only safeguarding.

Results against GCG text attacks. We also evaluate the robustness of Llama Guard 3 Vision against GCG, a strong attack that optimizes a text suffix that when appended to model inputs, maximizing the probability of predicting *safe*, similar to PGD. When Llama Guard 3 Vision is applied to classify the safety of a prompt, we find that GCG can circumvent the safety classifier the majority of the time (72% misclassified as safe). On the other hand, Llama Guard 3 Vision is reasonably robust in response classification, assuming that the attacker is only allowed to modify the prompt (30% vs. 16%). For a more successful attack, the adversary needs to place the GCG suffix in the agent responses, in which case the conversations misclassified as safe go up to 75%. However, in practice, this latter attack is much more challenging, since the attacker needs to both bypass the protections of the underlying language model and force it to output a previously-optimized suffix without knowledge of the model response,² or find a *universal* suffix that evades detection given arbitrary agent responses.

Recommendations to enhance robustness: we recommend enabling both prompt and response classification. In particular, against GCG we recommend the use of perplexity filters to detect adversarial inputs (Jain et al., 2023), although such filters can likely be circumvented with more sophisticated attack objectives.

5 Related Work

To prevent LLMs from producing harmful outputs, current mitigation strategies can be broadly categorized into two types: *model-level* and *system-level*. Model-level mitigation techniques include safety alignment methods such as instruction-tuning and RLHF (Wei et al., 2021; Ouyang et al., 2022), which aim to align the behavior of LLMs with human values. Recent works (Qi et al., 2024; Zou et al., 2024; Zhang et al., 2024) have also focused on robustifying safety alignment against adversarial attacks at the model level. System-level mitigation involves combining external components, such as safety classifiers (Inan et al., 2023; Llama Team, 2024b; Zeng et al., 2024), with base models (i.e., chatbots and agents). These safety classifiers can be used for prompt classification, response classification, or both. However, most system-level safety classifiers for LLM input-output safeguard are text-only. As multimodal LLMs become increasingly popular, Llama Guard 3 Vision serves as a starting point and baseline to encourage practitioners and researchers to further develop and adapt them to meet the evolving needs for multimodal LLM safety.

6 Limitations & Broader Impacts

There are some limitations associated with Llama Guard 3 Vision. First, Llama Guard 3 Vision itself is an LLM fine-tuned on Llama 3.2-vision. Thus, its performance (e.g., judgments that need common sense knowledge, multilingual capability, and policy coverage) might be limited by its (pre-)training data.

Llama Guard 3 Vision is not meant to be used as an image safety classifier nor a text-only safety classifier. Its task is to classify the multimodal prompt by itself, or the multimodal prompt along with the text response. It was optimized for the English language and only supports one image at the moment. Images will be rescaled into 4 chunks each of 560×560 , so the classification performance may vary depending on the actual image size. For text-only mitigation, we recommend using other safeguards in the Llama Guard family of models, such as Llama Guard 3-8B³ or Llama Guard 3-1B⁴, depending on your use case.

Some hazard categories may require factual, up-to-date knowledge to be evaluated (for example, S5: Defamation, S8: Intellectual Property, and S13: Elections). We believe that more complex systems should be deployed to accurately moderate these categories for use cases highly sensitive to these types of hazards, but Llama Guard 3 Vision provides a good baseline for generic use cases.

Lastly, as an LLM, Llama Guard 3 Vision could be susceptible to adversarial attacks (Zou et al., 2023; Carlini et al., 2024; Madry, 2017) that could bypass or alter its intended use. As shown in Section 4.2, in which we deploy Llama Guard 3 Vision for multimodal LLM output-only filtering under a realistic threat model scenario, the effectiveness of adversarial attacks can be mitigated to some extent. In practice, we suggest

²For a worst-case analysis, we optimized GCG suffixes placed after the harmful response it is evaluated on. In practice, the adversary may not have access to agent responses during optimization unless they have access to the underlying language model. ³https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/8B/MODEL_CARD.md

⁴https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL_CARD.md

combining Llama Guard 3 Vision with robust model safety alignment (Qi et al., 2024; Zhang et al., 2024) and other system components (e.g., a perplexity filter) for better performance against malicious users.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.
- Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. arXiv preprint arXiv:2404.09932, 2024.
- Tom B Brown. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. arXiv preprint arXiv:2305.17126, 2023.
- Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? Advances in Neural Information Processing Systems, 36, 2024.
- Google Gemini Team. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.
- Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. arXiv preprint arXiv:2404.05993, 2024.
- Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv preprint arXiv:2312.06674, 2023.
- Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. arXiv preprint arXiv:2309.00614, 2023.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023.
- AI @ Meta Llama Team. The llama 3 herd of models, 2024a. https://arxiv.org/abs/2407.21783.
- AI @ Meta Llama Team. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/ Llama-Guard2/MODEL_CARD.md, 2024b.
- Aleksander Madry. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021.
- OpenAI. Openai ol system card, 2024. https://cdn.openai.com/ol-system-card-20240917.pdf.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 2022.

- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings* of the IEEE international conference on computer vision, pages 2641–2649, 2015.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.
- Javier Rando, Hannah Korevaar, Erik Brinkman, Ivan Evtimov, and Florian Tramèr. Gradient-based jailbreak images for multimodal fusion models. arXiv preprint arXiv:2410.03489, 2024.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36, 2024.
- C Szegedy. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.
- Florian Tramer. Detecting adversarial examples is (Nearly) as hard as classifying them. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 21692–21702. PMLR, 17–23 Jul 2022.
- Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. arXiv preprint arXiv:2404.12241, 2024.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems, 36, 2024.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, 2021.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.
- Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. Rigorllm: Resilient guardrails for large language models against undesired content. arXiv preprint arXiv:2403.13031, 2024.
- Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. arXiv preprint arXiv:2407.21772, 2024.
- Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel M Bikel, Jason Weston, and Eric Michael Smith. Backtracking improves generation safety. arXiv preprint arXiv:2409.14586, 2024.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043, 2023.
- Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with short circuiting. arXiv preprint arXiv:2406.04313, 2024.

Appendix

A Acknowledgement

This work was made possible by a large group of contributors. We extend our gratitude to the following people: Tamar Glaser, Govind Thattai, Jana Vranes, Saghar Hosseini, Guillem Braso, Madian Khabsa, Archie Sravankumar, Ning Zhang, Ankit Ramchandani, Beto de Paola, Varun Vontimitta, Vincent Gonguet, Joe Spisak, Rachad Alao.

B PGD and GCG implementation details

For the PGD attack, we follow Madry (2017) and take steps on the image pixels in directions following signs of the gradient (computed to maximize the log likelihood of a safe classification). In our implementation, we set step size $\alpha = 0.1$, and take at most 100 iterations and early stop when the attack is successful. Images are projected back into the $l_{\infty} = \epsilon$ ball after each iteration. For the GCG attack, we take an off-the-shelf implementation (NanoGCG; Zou et al., 2023). During optimization, we run the attack for up to 100 steps (most successful attacks converge with fewer than 50 steps), using a search width of 64 and considering top 32 token replacements. Similar to GCG, we early stop when the attack is successful.